(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 December 2001 (13.12.2001)

PCT

(10) International Publication Number
WO 01/95312 A1

(51) International Patent Classification[7]: G10L 15/26

(21) International Application Number: PCT/IB01/00755

(22) International Filing Date: 2 May 2001 (02.05.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/590,708 8 June 2000 (08.06.2000) US

(71) Applicant: **NOKIA MOBILE PHONES LTD.** [FI/FI]; Keilalahdentie 4, FIN-02150 Espoo (FI).

(71) Applicant *(for LC only)*: **NOKIA INC.** [US/US]; 6000 Connection Drive, Irving, TX 75039 (US).

(72) Inventor: **HARIHARAN, Ramalingam**; Lindforsinkatu 6 A 12, FIN-33721 Tampere (FI).

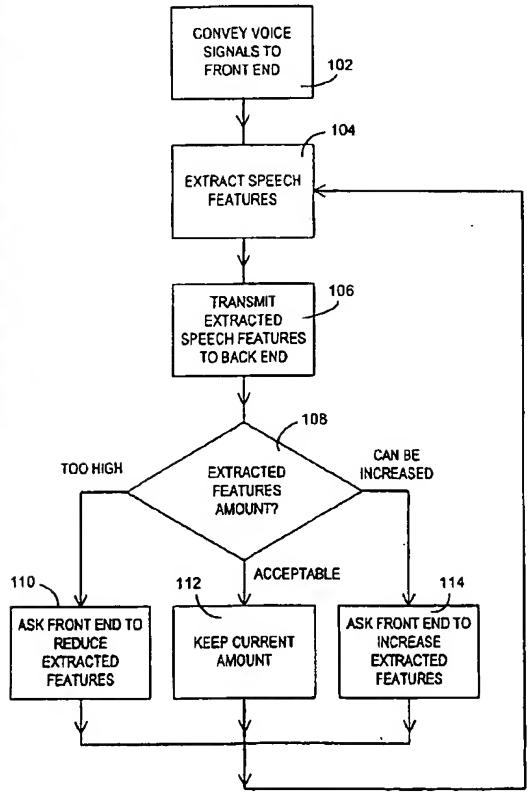
(74) Agent: **MAGUIRE, Francis, J.**; Ware, Fressola, Van Der Sluys & Adolphson LLP, 755 Main Street, P.O. Box 224, Monroe, CT 06468 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE,

[Continued on next page]

(54) Title: METHOD AND SYSTEM FOR ADAPTIVE DISTRIBUTED SPEECH RECOGNITION



(57) Abstract: A method and system for distributed speech recognition in a communications network which includes at least one speech processing server coupled to a plurality of communication terminals, wherein a front-end device residing at each terminal is used to extract an amount of speech features from voice signals and a back-end device residing at the server to recognize words from the extracted features. The front-end device is configurable so that the amount of speech features can be adjusted based on the prevailing conditions that affect the speech processing. At the server side, a traffic monitoring device is used to determine how many users are accessing the back-end recognizer so that the amount of extracted features can be increased when the channel traffic is light and reduced when the traffic is heavy.

WO 01/95312 A1

IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:

— *with international search report*

METHOD AND SYSTEM FOR ADAPTIVE DISTRIBUTED SPEECH RECOGNITION

Field of the Invention

The present invention relates generally to the field of speech recognition and, more particularly, to distributed speech recognition systems and methodology.

Background of the Invention

Speech recognition technology allows a user of a communications network to access computer services without using a keyboard to type in words, while a spoken language system provides user-computer interaction which enables natural conversations between people and machines. In particular, Distributed Speech Recognition (DSR) systems allow a user to give a verbal command, or dictate a memo to a speech processing device at one location and have the spoken words converted into written texts by a speech recognizer at another location. For example, the user can speak into a wireless device such as a mobile phone but the voice is recovered by a network device at a remote location. One of the emerging applications of DSR is a Voice Browser or a Wireless Application Protocol (WAP) Browser which allows anyone who has a telephone to access Internet-based services without being near a computer. DSR has many benefits. For example, voice interaction eliminates the need of having a keypad on a mobile device where physical space is limited for keypads and displays.

A DSR system is roughly divided into a front-end portion and a back-end portion. The front-end algorithm converts the input speech waveform signal into feature parameters which provide a compact representation of input speech, while retaining the information essential for speech recognition. The back-end algorithm performs the actual recognition task, taking feature parameters as input and performing a template-

matching operation to compare the features with reference templates of the possible words to be recognized.

In traditional Automatic Speech Recognition (ASR), both the front end and back end are located at the speech recognition server which is accessed through the Public Switched Telephone Network (PSTN) speech connection. If the speech signal comes from a mobile phone user, significant degradation of speech recognition accuracy may result from speech coding inaccuracies and radio transmission errors. Moreover, if the recognition results from ASR are used to drive a service which returns data to the user terminal, separate speech and data connections between the user terminal and the service are required.

DSR solves these problems of ASR by placing the front-end at the user terminal and transmitting feature parameters instead of the encoded speech waveform to the ASR server. Usually, feature parameters require less bandwidth for radio transmission than the encoded speech waveform. The feature parameters can, therefore, be sent to the ASR server using a data channel. This will eliminate the need for a high bit rate speech channel. Moreover, a low rate data transmission is less affected by noise and distortion, as compared to a speech channel transmission. Furthermore, if the data channel is equipped with error correction coding, the radio interface errors are no longer an issue. The full duplex data connection used to transmit the features to the ASR server can be used to send the response data (or the encoded speech) from the ASR server to the user terminal.

While DSR solves the problems with reduced ASR recognition accuracy and requires only one data connection for speech and data, it has the disadvantage that a standardized algorithm must exist for calculating feature parameters. The European Telecommunications Standard Institute (ETSI) is currently in the process of establishing the standard for DSR signal processing. ETSI has published in ETSI ES 201 108

V1.1.2 a standard algorithm for front-end feature extraction and their transmission. The standard algorithm calculates feature vectors with fourteen components for each 10ms frame of speech. In particular, this ETSI publication covers the algorithm for front-end feature extraction to create Mel-Frequency Cepstral Coefficients (MFCC).

Another disadvantage of the present DSR methodology is that the ASR server must be able to receive and use the features coming from the standard front end. Therefore, to support DSR, ASR vendors will have to modify their ASR engines to accommodate the DSR features. Depending on the technology used, this may be a minor undertaking or a technical challenge. If the feature vectors are sent to the ASR server using the fourteen components for each 10ms frame of speech, the resulting bit rate would be 44.8 kbps, assuming floating point coefficients and no framing overhead. This bit rate is clearly too high for cellular data channels. For this reason, the ETSI standard also includes a feature compression algorithm to provide an efficient way to transmit the coefficients in a lower data transmission rate. This compression algorithm combines 24 feature vectors, each of which is calculated from one 10ms frame of speech, to a multiframe of 143 bytes. This yields a bit rate of roughly 4767 bps. The ETSI publication also includes the formatting of the extracted features with error protection into a bitstream for transmissions, and the decoding of the bitstream to generate the front-end features at a back-end receiver together with the associated algorithm for channel error mitigation. Nokia ETSI-STQ WI008 also discloses a front-end algorithm for feature vector extraction. Cepstrum is a term for the inverse Fourier Transform of the logarithm of the power spectrum of a signal, and mel-frequency warping is a process of non-linearly modifying the scale of the Fourier transform representation of the spectrum. From the mel-frequency warped Fourier transform representation of the log-

magnitude spectrum, a set of cepstral coefficients or parameters are calculated to represent the speech signals. The extracted cepstral coefficients or parameters are known as feature vectors. They are conveyed to the back-end recognizer to perform the actual probability estimation and classification in order to reconstruct the spoken words. Because different speakers have different voices, talking speeds, accents and other factors that can affect a speech recognition system, it is important to have a good quality of feature vectors to ensure a good performance in speech recognition. Furthermore, environmental noises and distortion can also deteriorate the quality of feature vectors and influence the performance of the speech recognition system.

U.S. Patent No. 5,956,683, discloses a DSR system wherein extracted features are transmitted from a portable phone to a central communication center which has a word decoder for determining a linguistic estimate of the speech from the extracted features and providing an action signal to a transmitter in the communication center. Under a control element of the communication center, the transmitter sends estimated words or a command signal to the portable phone. The estimated words or the command signal will be used for dialing a phone number, providing information to the display screen of the portable phone, or relaying messages from an answering machine.

The performance of a speech recognition system, in general, correlates significantly with the number of features extracted from the front end and processed by the back end. Therefore, it is desirable to increase the amount of features extracted in the front-end in order to increase the performance. However, this also increases the complexities in the back-end recognizer situated at the server side because the recognizer has to process all the received features. In particular, in a DSR system where a back-end network recognizer is used to process the speech data from a plurality

of terminals, the processing power and time of the back-end recognizer imposes a limit on the amount of extracted features that can be transmitted from each terminal to be simultaneously processed. In the existing DSR system, the amount of extracted features is fixed and is usually determined by the maximum number of terminals that will share a back-end recognizer. In such a system, the preset level of speech recognition performance is determined based on the worst case traffic condition at the back-end recognizer. As such, the highest attainable performance is not usually fully attained.

It is advantageous and desirable to provide a DSR system with improved performance such that the performance of the speech recognition system can be extended beyond the limit derived from a worst-case analysis.

Summary of the Invention

It is an object of the invention to provide an adaptive or scalable distributed speech recognition (DSR) method and system wherein the amount of the speech-related features extracted from a voice signal by a front-end device can be varied in accordance with the prevailing traffic conditions of the back-end recognizer in a network.

Accordingly, one aspect of the present invention is a method of distributed speech processing using a speech recognition system in a communications network. The network includes at least one speech processing server coupled to a plurality of communication terminals, wherein the server has a speech recognition system for recognizing words from spoken voice signals conveyed to the terminals by the users' voices. The method includes the steps of:

extracting speech features from the voice signals; and

transmitting the extracted speech features to the server in order to recognize words from the extracted speech features, wherein the amount of the extracted speech features

can be varied and determined by the server based on prevailing conditions of the server.

Another aspect of the present invention is a speech recognition apparatus to be used in a communications network having at least one server coupled to a plurality of communication terminals. The apparatus includes:

a first device to extract speech features from one or more voice signals conveyed to one or more corresponding terminals; and

a second device to recognize words from the extracted speech features, wherein the second device resides at the server and the first device resides in said one or more terminals, and wherein the first device is configurable so that the amount of the speech features extracted can be controlled by the server depending on the prevailing conditions of the server.

The third aspect of the present invention is a speech recognition system for a communications network having at least one server coupled to a plurality of communication terminals, wherein the system is used to recognize words from voice signals conveyed to the terminals by corresponding users, and includes:

front-end devices residing at the terminals to extract speech features from the voice signals;

a back-end device residing at the server to recognize words from the extracted speech features,

conveying devices to convey the extracted features from the front-end devices to the back-end device, and

a transmitter to convey signals from the server to the front-end devices specifying the amount of speech features to be extracted as determined by the server.

The present invention will become apparent upon reading the description taken in conjunction with Figures 1 to 3.

Brief Description of the Drawings

Figure 1 is a diagrammatic presentation of a communications network showing a plurality of wireless terminals coupled to a server so as to allow a plurality of front-end devices to share a back-end device at the server.

Figure 2 is a block diagram illustrating a DSR (Distributed Speech Recognition) system having a front-end portion and a back-end portion.

Figure 3 is a flow chart illustrating the method of DSR, according to the present invention.

Best Mode for Carrying Out the Invention

Figure 1 shows a communications network **10** having a plurality of terminals **12** coupled to a server **18**. The terminals **12** can be mobile phones or other wireless or wired devices. Each of the terminals **12** has a voice input device such as a microphone **16** to allow a user to give verbal commands or input spoken messages to the terminals **12**. The spoken voice of the speaker is converted into voice signals by the microphone **16**. Each of the terminals **12** includes a front-end processor **20** responsive to a voice signal from such a microphone **16** to extract feature vectors and/or other speech-related information from the voice signals. The extracted features are conveyed as speech data **22** to a back-end speech recognizer device **40** situated at the server side. The front-end processor **20** is configurable such that it is capable of producing speech data with different complexities on demand. For example, the front-end processor **20** can be configured at anytime to extract more or less speech-related features from a speech frame. When the speech recognizer **40** is simultaneously used to process the speech data from a large number of terminals **12**, the speech recognizer **40** may require the terminals **12** to reduce the complexity in the feature extraction in order to accommodate the channel traffic condition at the speech recognizer **40**. Accordingly, the

server 18 sends a signal containing control data 24 to the terminals 12 indicating the optimal amount of feature vectors under the prevailing traffic condition. Upon receiving control data 24 from the server 18, the terminals 12 adjust the extracted amount of speech features as required. But when the channel traffic is light, the speech recognizer 40 can process speech data with higher complexity in order to improve the DSR performance. Accordingly, the speech recognizer 40 sends a new set of control data 24 to the engaging terminals 12 to specify the amount of the speech features to be extracted. This adaptive or scalable DSR system can help the user to obtain better speech recognition performance at the cost of higher complexity when there is a smaller number of terminals 12 simultaneously accessing the speech recognizer 40 at the server 18. Another criterion that could be used to determine the amount of extracted features is the prevailing environmental condition or signal-to-noise ratio (SNR) at the terminal side. It is understood that a cleaner environment requires a smaller set of extracted features without unduly compromising the speech recognition performance. A smaller set of feature vectors can lower the overall complexity of the recognition process and hence the computation time in the back-end portion.

Figure 2 shows the front-end device 20 and the back-end device 40 of the adaptive DSR apparatus of the present invention. As shown, the front-end device 20 includes a receiver 30 to receive voice signals from the voice input device 16, and a front-end speech processor 32 to extract feature vectors from the voice signals and to convey the extracted features to a data channel 34. The speech-related extracted features are conveyed as speech data 22 to the back-end device 40 for further processing. The back-end device 40 includes a speech data receiver 42 to receive speech data 22 from a plurality of terminals 12, and a back-end speech recognizer 44 which converts speech data 22 into words or

texts 46. The back-end device 40 further includes a traffic monitoring device 48 to determine the workload of the back-end speech recognizer 44 according to the number of simultaneous users and the complexity of the speech data 22 as received from the engaging terminals 12. The traffic monitoring device 48 provides a signal indication of the current workload of the back-end speech recognizer 44 to a decider 49. The decider can decide that the back-end has reached its capacity, even though more users are trying to access the back-end device 40 for speech processing services. Accordingly, decider 49 determines a reduced amount of features to be extracted by the front-end processor 32 under the prevailing traffic condition. The deduced amount of extracted features is then signaled by the decider to a control data device 50 which conveys control data 24 to the front-end device 20. The control data 24 can be conveyed in the form of control bits or any suitable form.

Similarly, when the channel traffic on the server side is low, it is possible to increase the amount of extracted features in order to improve the speech recognition performance, especially when the prevailing environmental noise conditions at the terminal side may affect the recognition task. Accordingly, a suitable amount of extracted features is conveyed as the control data 24 to the front-end device 20 of the engaging terminals 12. At the front-end side, a receiving device 36 conveys the received control data 24 to the front-end speech processor 32 so that the amount of extract features can be adjusted as required.

The method of the adaptive DSR, according to the present invention, is illustrated in Figure 3. As shown, when one or more users speak to the terminals 12 to ask the server 18 to perform a certain task, the speech or voice signals are conveyed to the front-end device 20 at step 102. According to the amount of extracted features currently required by the back-end device 40, the front-end device 20 extracts a set of feature vectors from the voice signals at step 104. The

extracted features are transmitted to the back-end device for speech recognition at step **106**. Based on the prevailing conditions, such as the channel traffic conditions at the server side and environmental noise conditions at the user side, the amount of extracted features are accessed at step **108**. If it is decided that the amount of extracted features is too high, the terminals **12** are asked to reduce the extracted amount at step **110**. If the extracted amount is acceptable, the terminals **12** are asked to keep the current extracted amount at step **112**. However, this step is optional because it does not affect the outcome of the process. If the prevailing conditions allow for an increased amount of extracted features, the terminals **12** are asked to adjust the amount accordingly at step **114**. The command signal specifying the required amount for feature extraction is sent to the front-end device **20** as the process loops back to step **104**.

Thus, the method, the device and the system for the adaptive distributed speech recognition, according to the present invention, have been disclosed in the preferred embodiments thereof. It will be understood by those skilled in the art that the foregoing and various other changes, omissions and deviations in the form and detail thereof may be made without departing from the spirit and scope of this invention. For example, the most commonly used approach for carrying out the feature extraction is the cepstral approach and the extracted feature vectors are the mel-frequency cepstral coefficients. However, the method, device and system for adaptive speech recognition of the present invention are not limited to the cepstral approach. They are equally applied to any other speech recognition approach. Furthermore, in the cepstral approach, the complexity of the front-end feature extraction depends on how the speech signals are sampled, how the Fast Fourier Transform is carried out, how the transformed spectrum is processed and how the mel-frequency cepstrum coefficients are computed. It is possible

to increase or decrease the amount of extracted feature vectors by adjusting the various processing steps in the cepstrum approach. It is also possible to adjust the complexity of the feature extraction by replacing the cepstrum approach with another different approach.

The communication terminals as used hereinabove to explain the present invention include mobile phones, communicators and other hand-held devices. However, the terminals can be any communication devices that are physically separated from the server such that they require a distributed speech recognition system for speech signal processing. In a narrow sense, the server is an Automatic Speech Recognition (ASR) server. In a broad sense, it can be generally referred to as a central communication center. The terminals can be coupled to the server via a radio link, but they can also be linked to the server through a different medium.

Therefore, the embodiments and method described hereinabove should be considered illustrative, but not restricting. The possibilities of implementing and using the invention are only restricted by the appended claims. Consequently, the various options of implementing the invention as determined by the claims, including the equivalent implementations, also belong to the scope of the present invention.

What is claimed is:

1. A method for distributed speech recognition to be used in a communications network having at least one server coupled to a plurality of terminals, wherein the speech recognition method is used to recognize words from voice signals conveyed to the terminals, said method comprising the steps of:

1) extracting speech features from the voice signals; and

2) transmitting the extracted speech features to the server in order to recognize words therefrom, wherein the amount of the extracted speech features can be adjusted based on prevailing conditions affecting speech recognition.

2. The method of claim 1, wherein the terminals have means to extract the speech features and the server has means to determine the amount of speech features to be extracted by the terminals based on the prevailing conditions, said method further comprising the step of:

3) conveying a command signal from the server to the terminals in order to adjust the amount of speech features to be extracted by the terminals.

3. The method of claim 2, wherein the command signal includes control bits indicative of the adjusted amount of extracted speech features.

4. The method of claim 1, wherein the prevailing conditions are determined by the number of terminals simultaneously transmitting the extracted speech feature to the server for speech recognition.

5. The method of claim 1, wherein the prevailing conditions are determined by environment noise conditions at the terminals.

6. The method of claim 1, wherein the extracted speech features include cepstral coefficients.

7. The method of claim 1, wherein the extracted speech features include mel-frequency cepstral coefficients.

8. The method of claim 1, wherein the terminals include mobile phones, each of which is coupled to the server via a radio link.

9. The method of claim 1, wherein the extracted speech features are transmitted to the server via a data channel.

10. A distributed speech recognition apparatus to be used for speech recognition processing in a network having at least one server coupled to a plurality of terminals, said apparatus comprising:

a front-end device in one or more of the terminals to extract speech features from voice signals conveyed to the terminals; and

a back-end device situated at the server to recognize words from the extracted speech features, wherein the front-end device is configurable so that the amount of the speech features extracted by the front-end device can be controlled by the server depending on prevailing conditions affecting the speech recognition processing.

11. The apparatus of claim 10, further comprising means for converting a voice of a user to voice signals.

12. The apparatus of claim 11, wherein the voice converting means comprise a microphone.

13. The apparatus of claim 10, wherein the terminals comprises one or more mobile phones, each of which is coupled to the server via a radio link.

14. The apparatus of claim 10, where the front-end device comprises means for transmitting the extracted speech feature to the back-end device.

15. The apparatus of claim 10, wherein the back-end device comprises means for monitoring the prevailing conditions.

16. The apparatus of claim 10, wherein the prevailing conditions include channel traffic at the back-end device.

17. The apparatus of claim 10, wherein the back-end device is capable of sending a signal to the front-end device in order to adjust the amount of speech features to be extracted by the front-end device.

18. A distributed speech recognition system to be used in a network having at least one server coupled to a plurality of terminals, wherein the system is used to recognize words from voice signals conveyed to the terminals, said system comprising:

a front-end device residing at each of the terminals for extracting an amount of speech features from the voice signal, wherein the amount of extracted speech features is adjustable;

a back-end device residing at the server to recognize words from the extracted speech features;

means for conveying the extracted features from the front-end means to the back-end means, and

means for conveying a command signal from the server to the front-end device indicating the amount of speech features to be extracted by the front-end means.

19. The system of claim 18, wherein the amount of speech features to be extracted as indicated in the command signal is determined by the server based on prevailing conditions affecting the back-end device.

20. The system of claim 19, wherein the prevailing conditions include channel traffic conditions related to the number of terminals simultaneously accessing the back-end device for speech recognition purposes.

21. The system of claim 18, wherein the amount of speech features to be extracted as indicated in the command signal is determined by the server based on prevailing environmental noise conditions affecting the front-end device.

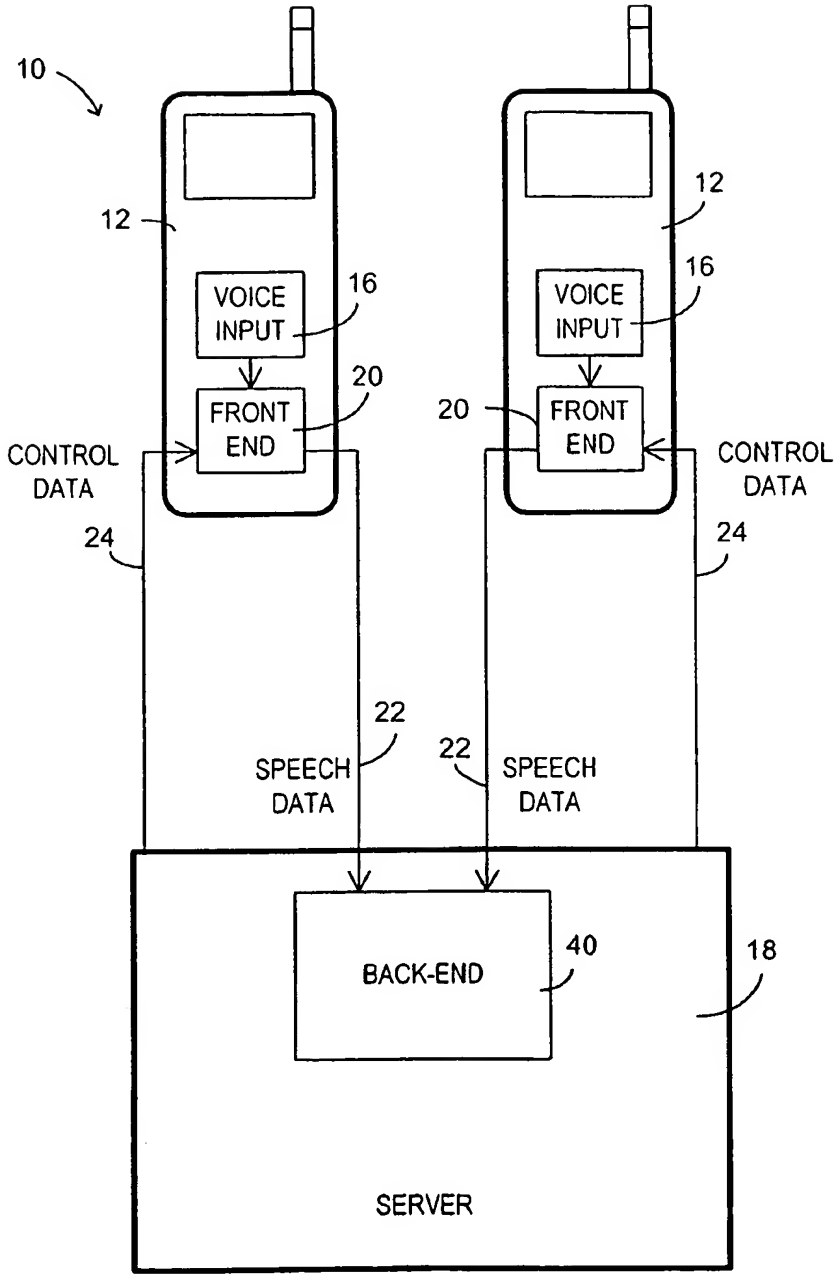

10
12
VOICE INPUT
16
20
FRONT END
CONTROL DATA
24
22
SPEECH DATA
12
VOICE INPUT
16
20
FRONT END
CONTROL DATA
24
22
SPEECH DATA
BACK-END
40
18
SERVER

FIG. 1

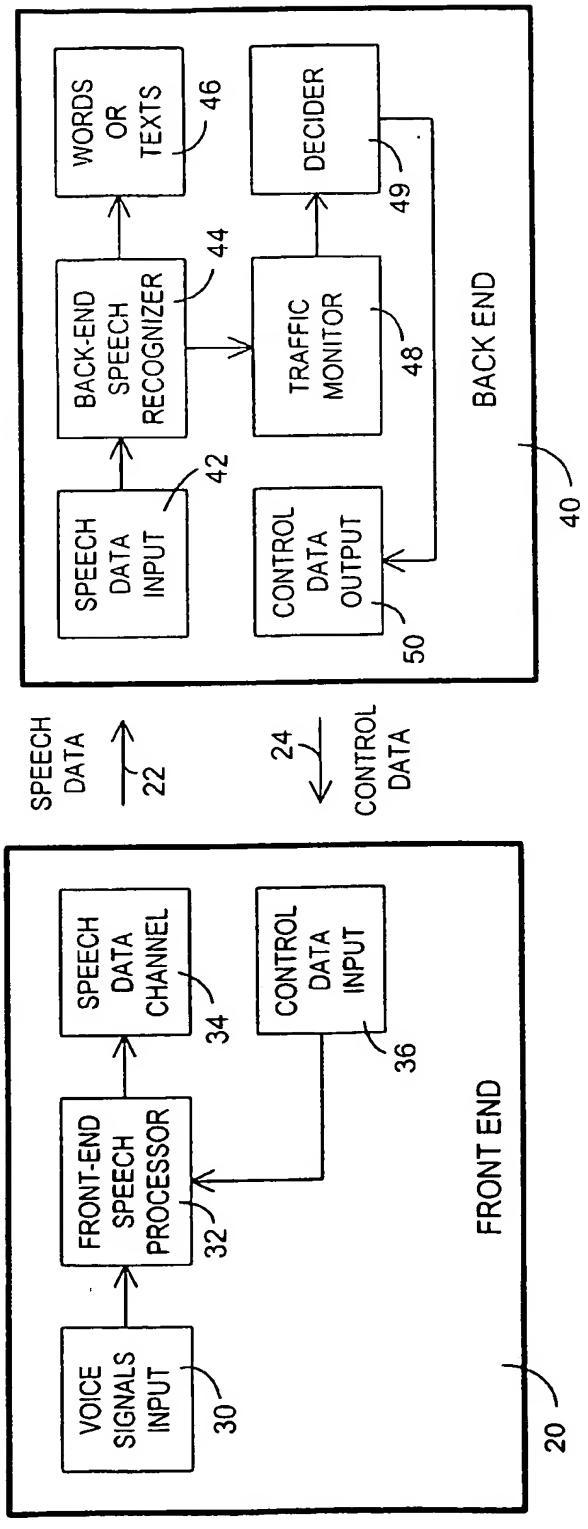

FRONT END
20
VOICE SIGNALS INPUT
30
FRONT-END SPEECH PROCESSOR
32
SPEECH DATA CHANNEL
34
CONTROL DATA INPUT
36
SPEECH DATA
22
24
CONTROL DATA
BACK END
40
SPEECH DATA INPUT
42
BACK-END SPEECH RECOGNIZER
44
WORDS OR TEXTS
46
TRAFFIC MONITOR
48
DECIDER
49
CONTROL DATA OUTPUT
50

FIG. 2

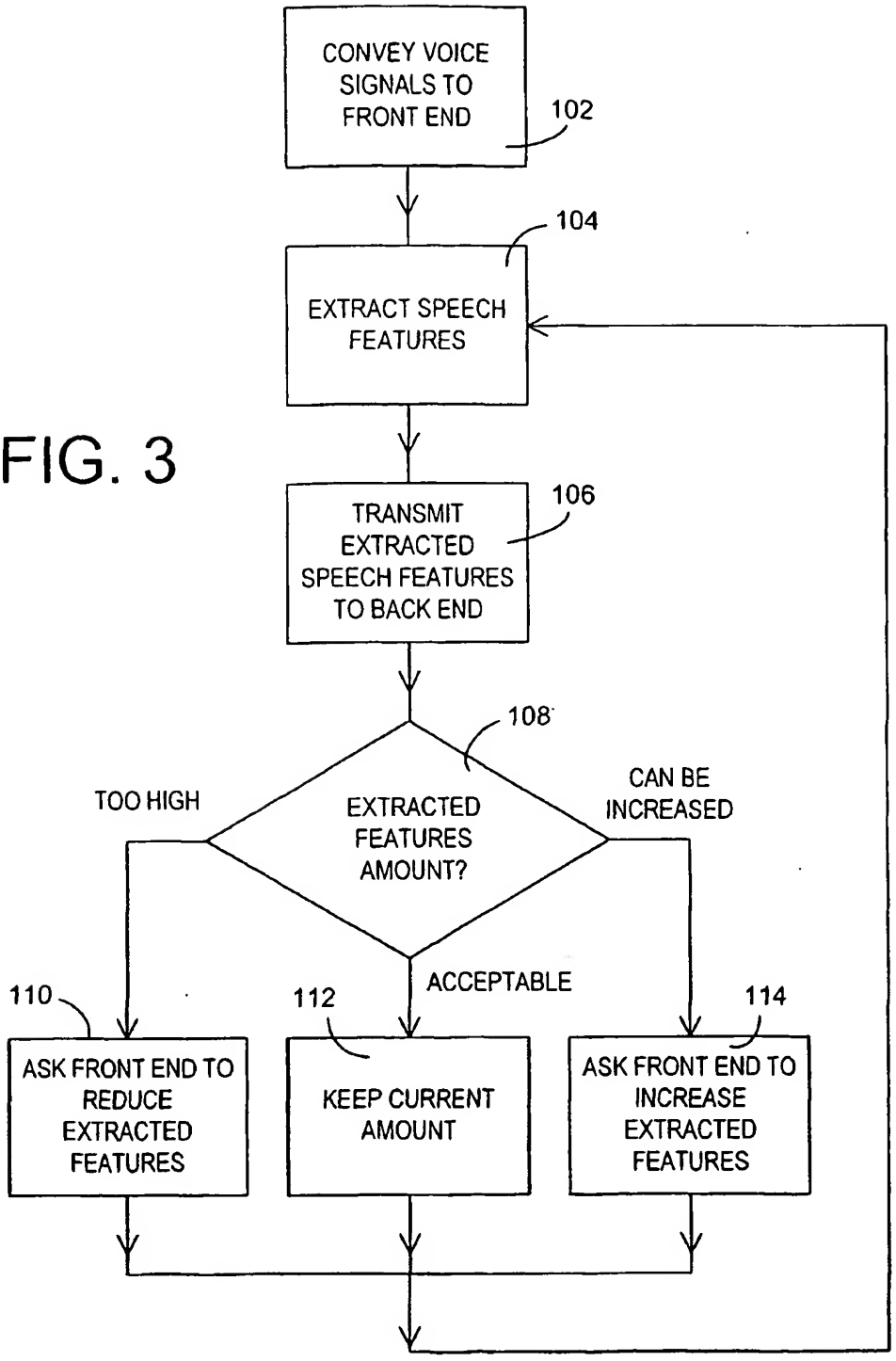

CONVEY VOICE SIGNALS TO FRONT END
102
104
EXTRACT SPEECH FEATURES
TRANSMIT EXTRACTED SPEECH FEATURES TO BACK END
106
108
EXTRACTED FEATURES AMOUNT?
TOO HIGH
CAN BE INCREASED
ACCEPTABLE
110
112
114
ASK FRONT END TO REDUCE EXTRACTED FEATURES
KEEP CURRENT AMOUNT
ASK FRONT END TO INCREASE EXTRACTED FEATURES

FIG. 3